Discussion Papers No. 244, January 1999
Statistics Norway, Research Department

Joe Sexton and Anders Rygh Swensen

ECM-algorithms that converge at the rate of EM

Abstract:
This paper describes a way of constructing an ECM algorithm such that it converges at the rate of the EM algorithm. The approach is motivated by the well known conjugate directions algorithm, and a special case of it is when the parameters corresponding to different CM steps are orthogonal. Three examples are given illustrating the approach. Possible implications of the theme for the ECME algorithm are briefly discussed.

Keywords: EM algorithm, ECM algorithm, ECME algorithm, missing data, conjugate directions algorithm, orthogonal parameters, rate of convergence.

JEL classification: C63, C24

Address: Joe Sexton, Statistics Norway, Research Department, Box 8131 Dep. N-0033 Oslo, Norway. E-mail: joe.sexton@ssb.no

Anders Rygh Swensen, Statistics Norway, Research Department, Box 8131 Dep. N-0033 Oslo, Norway. E-mail: anders.rygh.swensen@ssb.no

Discussion Papers comprise research papers intended for international journals or books. As a preprint a Discussion Paper can be longer and more elaborate than a standard journal article by including intermediate calculation and background material etc.

Abstracts with downloadable PDF files of
Discussion Papers are available on the Internet: http://www.ssb.no

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
P.O. Box 1260
N-2201 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no

1 Introduction

The EM-algorithm introduced by Dempster, Laird and Rubin (1977) (hereafter DLR) is an elegant and popular algorithm for finding maximum likelihood estimates in missing data situations. However, in some of these situations , in particular those where the complete data likelihood is complicated and must be maximized numerically, the EM may be less tractable. This is because its implementation leads to nested iterations and thereby a possibly unstable algorithm. To better handle such situations Meng and Rubin (1993) (hereafter MR) introduced a closely related algorithm they called the ECM-algorithm. Here the M-step of the EM is replaced by a sequence of conditional maximization (CM-) steps. The motivation being that although the complete data likelihood itself may require numerical iteration, the maximization over subvectors of the parameter vector are often, conditionally given the value of the other parameters, in closed form. And even when this is not the case, reducing the dimension of the numerical optimization, increases stability of the algorithm.

The ECM can in some situations lead to substantially simpler algorithms compared with the EM. A question one might ask however, is how much this added simplicity and stability has cost measured by a slower convergence rate of the algorithm. Obviously, the price will vary across situations, and some times the ECM may even converge at a faster rate than EM, see Meng (1994). Such situations are however, as MR point out, not typical in practice, but a complete characterization of these seems difficult.

There has been suggested a number of different ways of speeding up the convergence of the EM algorithm. A brief review of these is given in Meng and van Dyke (1997) where they also propose another possible approach to this problem. Here, however, we discuss a way of constructing the CM steps such that the resulting ECM algorithm converges at the same rate as EM. The approach is motivated by the well known conjugate-directions algorithm for function optimization, see Luenberger (1989) or Zangwill (1969). A special and important case of the situation we discuss is when the parameters corresponding different CM steps are orthogonal (i.e their maximum likelihood estimators are asymptotically uncorrelated), here the ECM will in large samples converge at the same rate as EM, and thus the added simplicity and stability of ECM over EM is basically free of charge. Cox and Reid (1987) note, referring to complete data situations, that orthogonal parameters may simplify the numerical maximization of the likelihood.

The remainder of this paper is as follows. In section 2 the EM and ECM algorithms are defined along with the measuring of convergence rate. In section 3 the main result is stated, and examples are given in section 4. Section 5 discusses briefly, through an example, the possible implications our underlying theme has for the ECME algorithm, a close relative of the ECM. Concluding remarks are given in the sixth and final section.

2 Background material

2.1 The EM and ECM algorithms

Missing data often complicates the likelihood function and makes it difficult to manipulate analytically. To see why, let $Y_{COM} = (Y_{OBS}, Y_{MIS})$ be the complete data, where Y_{OBS} denotes the observed data and Y_{MIS} the missing data. Further let $\lambda \in R^p$ be the parameter vector, and f be the complete data density. The likelihood of the observed data is then:

$$L_{OBS}(\lambda) = \log \int_{Y_{MIS}} f(Y_{COM}; \lambda) dY_{MIS},$$

and it is this integrating out of the missing data that complicates L_{OBS}. The ECM, and thus the EM which is presented as a special case of the ECM, maximizes L_{OBS} via the following procedure.

The ECM generates a sequence of parameter values by, given $\lambda^{(0)}$, iterating the steps:
E-step

Compute:

$$Q(\lambda|\lambda^{(t)}) = E(L_{COM}(\lambda)|Y_{OBS};\lambda^{(t)}) \tag{1}$$

CM-steps

For each $s = 1, .., S$ find $\lambda^{(t+s/S)}$ such that:

$$Q(\lambda^{(t+s/S)}|\lambda^{(t)}) = max_{\lambda}Q(\lambda|\lambda^{(t)}) \tag{2}$$

under the constraint $g_s(\lambda) = g_s(\lambda^{(t+(s-1)/S)})$, where $G = (g_s(\lambda); s = 1, ..S)$ are preselected vector functions.
The parameter-sequence $(\lambda^{(t)})_0^{\infty}$ generated by this algorithm has (under regularity conditions see MR for the ECM, and DLR and Wu (1983) for the EM) the following two very appealing properties:
1)

$$L_{OBS}(\lambda^{(t+1)}) \geq L_{OBS}(\lambda^{(t)})$$

2)

$$\lim_{t\to\infty} DL_{OBS}(\lambda^{(t)}) = 0,$$

here D denotes the differential operator.

The EM comes about by choosing:

$$g(\lambda) = \text{a constant}, \tag{3}$$

and thus the one and only CM step consists in maximizing Q over the entire parameter space. Besides (3) i.e the EM algorithm, the most frequently occurring choice of the G functions are:

$$g_s(\lambda) = (\lambda_1, .., \lambda_{s-1}, \lambda_{s+1}, .., \lambda_S) \tag{4}$$

This implies that the s-th CM-step consists of maximizing the Q-function over the subvector λ_s while holding the remaining elements of the parameter vector fixed. This subclass of the ECM-algorithms is called, Meng and Rubin (1992), PECM-algorithm, with the P meaning 'partitioned'. Without missing data the ECM is a special case of the cyclic coordinate ascent method for function maximization, see Zangwill (1969).

2.2 The rate of convergence

Here we follow the setup in Meng (1994). Any iterative estimation algorithm implicitly defines a mapping 'M' from the parameter-space onto itself, such that $M(\lambda^{(t)}) = \lambda^{(t+1)}$. Supposing M is differentiable and that we are close enough to the limit point λ^* of $(\lambda^{(t)})_0^{\infty}$, we have, letting $DM()$ denote the Jacobian of the transformation M, that:

$$(\lambda^{(t+1)} - \lambda^*) = (\lambda^{(t)} - \lambda^*)DM(\lambda^*),$$

ignoring terms of higher order. The matrix $DM(\lambda^*)$ is often referred to as the matrix rate of convergence.

The observed rate of convergence is, reasonably, defined as:

$$r = \lim_{t\to\infty} \frac{\|\lambda^{(t+1)} - \lambda^*\|}{\|\lambda^{(t)} - \lambda^*\|},$$

which is (Meng (1994)) equal to the largest eigenvalue of $DM(\lambda^*)$. Note that a large rate implies slow convergence. The speed of the algorithm is defined as $s = 1 - r$.

DLR showed that for the mapping M^{EM} defined by the EM:

$$DM^{EM}(\lambda^*) = I_{MIS}(\lambda^*)I^{-1}_{COM}(\lambda^*) \tag{5}$$

where

$$I_{MIS}(\lambda^*) = -\int (D_{\lambda\lambda} \log f(Y_{MIS}|Y_{OBS};\lambda^*)) f(Y_{MIS}|Y_{OBS};\lambda^*) dY_{MIS}$$

and

$$I_{COM}(\lambda^*) = -\int (D_{\lambda\lambda} \log f(Y_{OBS}, Y_{MIS};\lambda^*)) f(Y_{MIS}|Y_{OBS};\lambda^*) dY_{MIS} \tag{6}$$

Here D represents the differential operator as before,D_λ differentiation with respect to λ, and $D_{\lambda\lambda} = D_\lambda D_\lambda$.

Meng (1994) showed that for the ECM-algorithm:

$$DM^{ECM}(\lambda^*) = DM^{EM}(\lambda^*) + (I_p - DM^{EM}(\lambda^*)) \prod_{s=1}^{S} P_s, \tag{7}$$

where

$$P_s = \nabla_s(\nabla_s^T I^{-1}_{COM}(\lambda^*)\nabla_s)^{-1}\nabla_s^T I^{-1}_{COM}(\lambda^*), \qquad s = 1,..,S \tag{8}$$

with $\nabla_s = Dg_s(\lambda^*)$.

3 Main result

In this section we examine ECM-algorithms where in each CM-step, say the s-th step following the t-th E-step, one maximizes $Q(\quad|\lambda^{(t)})$, defined in (1), over a set of vectors in the parameter space, denote these by $d_s = (d_s^{(1)} : .. : d_s^{(m_s)})$, where $d_s^{(i)} \quad i = 1,..,m_s$ are column vectors, that are constructed to have the property:

$$d_i^T I_{COM}(\lambda^*) d_j = 0 \qquad i \neq j \qquad i,j \in (1,..,S) \tag{9}$$

We say that such vectors are I_{COM}-orthogonal.

Now we proceed to show that an ECM-algorithm constructed in this manner will converge at the same rate as EM. First some observations.

Observation 1 *The span of d_s and the span of the column vectors of ∇_s (defined in (8)) are orthogonal complements.*

Proof: The orthogonality of the two sets of vectors follows from the fact that $g_s(\lambda)$ is held fixed when we move along the vectors of d_s. That the column vectors of ∇_s span all vectors orthogonal to d_s is a consequence of the fact that if they didn't there would be vectors not spanned by d_s but that satisfied $g_s(\lambda)$, which would contradict that d_s are the only directions being searched over in the s-th CM-step.

Observation 2 *The set of vectors* d_s *and* $I_{COM}d_i$ $i \neq s$ *span* R^p.

Proof: The column vectors of d_s are linear independent, and by construction orthogonal to the vectors $I_{COM}d_i$ $i \neq s$, and therefore also linear independent to this set. The vectors $I_{COM}d_i$ $i \neq s$ are linear independent because the set d_i $i \neq s$ is and I_{COM} is invertible, and thus we have a set of p linear independent vectors which necessarily span R^p.

Observation 3 *There exists an invertible matrix* α_s *such that:*

$$\nabla_s = I_{COM} D_s \alpha_s, \tag{10}$$

where $D_s = (d_1 : .. : d_{s-1} : d_{s+1} : .. : d_S)$.

Proof: That the column vectors of ∇_s and the column vectors of $I_{COM}D_s$ span the same space is a consequence of observation 1 and 2. Thus there must exist an invertible matrix relating the two sets of vectors, her denoted by α_s.

Now it is shown that the matrix rate of convergence of an ECM-algorithm that has been constructed as described in the beginning of this section, is identical to that of EM.

Proposition 1 *If the vectors that are searched over in each CM-step are* I_{COM}*-orthogonal to the search vectors in the other CM-steps then:*

$$DM^{ECM} = DM^{EM} \tag{11}$$

Proof: By observation 3 we have that:

$$\nabla_s = I_{COM} D_s \alpha_s$$

Consider (8):

$$\nabla_s^T I_{COM}^{-1} \nabla_s = \alpha_s^T D_s^T I_{COM} D_s \alpha_s = \alpha_s^T \alpha_s,$$

assuming, without loss of generality, that the search vectors have been normalized and orthogonalized, so that $D_s^T I_{COM} D_s = I$. Now:

$$\nabla_s (\nabla_s^T I_{COM}^{-1} \nabla_s)^{-1} \nabla_s^T = I_{COM} D_s \alpha_s (\alpha_s^T \alpha_s)^{-1} \alpha_s^T D_s^T I_{COM} = I_{COM} D_s D_s^T I_{COM},$$

where the last equality follows from α_s being invertible. Thus:

$$P_s = I_{COM} D_s D_s^T I_{COM} I_{COM}^{-1} = I_{COM} D_s D_s^T.$$

And then:

$$\prod_{s=1}^{S} P_s = \prod_{s=1}^{S} I_{COM} D_s D_s^T = I_{COM} D_1 (\prod_{s=1}^{S-1} D_s^T I_{COM} D_{s+1}) D_S^T.$$

Because of the I_{COM}-orthogonality of the search vectors corresponding different CM-steps, we have that:

$$\prod_{s=1}^{S-1} D_s^T I_{COM} D_{s+1} = 0,$$

which implies (11) through (7) .

Usually it is the case that we do not have any information about I_{COM} at the outset of the estimation. Thus the search vectors cannot be determined once and for all at the start of the algorithm. These vectors will need to be successively updated with each new element of the parameter sequence generated by the iterative algorithm.

A convenient by-product of using an ECM algorithm as discussed above, is that the SECM algorithm, the Supplemented-ECM algorithm (see Meng and Rubin (1992)), which uses the matrix rate of convergence of the ECM to compute the observed information matrix, is simplified. The SECM algorithm is the counterpart to the SEM algorithm introduced by Meng and Rubin (1991). Meng and Rubin (1992) define $DM^{CM} = \prod_{s=1}^{S} P_s$, where P_s is as in (8), and derive the SECM-algorithm from the relation

$$I_{OBS} = (I - DM^{ECM})(I - DM^{CM})^{-1} I_{COM},$$

where I_{OBS} is the observed information matrix, and I_{COM} as defined earlier. However in the situation discussed in Proposition 1 we have that: $DM^{CM} = 0$, giving the mentioned simplification of the SECM.

Although we do not know I_{COM}, we do in some situations know something about the structure of this matrix in large samples. For example, when some of the parameters are asymptotically orthogonal, i.e the corresponding elements of the information matrix are zero. This motivates the following proposition.

Proposition 2 *If the vectors that are searched over in each CM-step are asymptotically I_{COM}-orthogonal i.e:*

$$d_j^T \frac{1}{n} I_{COM}^{(n)} d_k \to^P 0 \qquad j \neq k \qquad as\ n \to \infty, \tag{12}$$

where n denotes the number of observations, and $I_{COM}^{(n)}$ the I_{COM} matrix with n observations evaluated at the maximum likelihood estimate derived form these observations. Then:

$$\| DM_{(n)}^{ECM} - DM_{(n)}^{EM} \| \to^P 0 \qquad as\ n \to \infty. \tag{13}$$

Proof: The proof follows the same lines as the proof for Proposition 1.

The most important, from a practical point of view, case of asymptotical I_{COM}-orthogonality occurs when $\frac{1}{n} I_{COM}^{(n)}$ converges to a block-diagonal matrix. Since it is reasonable to expect that $\frac{1}{n} I_{COM}^{(n)}$ approaches the expected information matrix in the complete data model, denoted by $i(\lambda)$, the block-diagonality of the limit of $\frac{1}{n} I_{COM}^{(n)}$ can be inferred from that of $i(\lambda)$, which by definition means that the parameters corresponding different blocks are orthogonal in the complete data situation. In such cases it is natural to let each d_s, for $s = 1, .., S$, consist of the subset of the standard basis vectors that span the rows corresponding the s-th block of $i(\lambda)$. This is in other words the ECM algorithm that in each CM step maximizes the Q-function in (1) over a subset of the parameter vector that is orthogonal, in the complete data situation, to the parameters being held fixed. If $\frac{1}{n} I_{COM}^{(n)}$ does not converge to a block diagonal matrix,

then we shall see that in some situations it is possible to reparameterize the model such that the reparameterized model does have this property.

Although $I_{COM}^{(n)}$ in (12) is not the observed information matrix in the complete data situation, but the expectation of this matrix, conditional on Y_{OBS}, it is reasonable to expect, as pointed out above, that $\frac{1}{n}I_{COM}^{(n)}$ converges, in some sense, to $i(\lambda)$. Thus, under appropriate conditions, the property $d_j^T i(\lambda) d_k = 0$ will imply (12). We now discuss briefly conditions under which this implication is true.

For example, by an application of the triangle inequality one can verify that the conditions:

$$\frac{1}{n}D_{\lambda\lambda}L_{COM}(\lambda_n) \rightarrow^{L^1} \lim_{n\to\infty} E(\frac{1}{n}D_{\lambda\lambda}L_{COM}(\lambda)) = -i(\lambda) \qquad \text{as } n \to \infty, \tag{14}$$

where $(\lambda_n)_{n=0}^{\infty}$ here is any sequence converging to the true parameter vector, here denoted by λ, and

$$|d_j^T \frac{1}{n}D^{20}Q(\lambda_n|\lambda_n)d_k - d_j^T \frac{1}{n}D^{20}Q(\lambda_n|\lambda)d_k| \rightarrow^P 0 \qquad \text{as } n \to \infty, \tag{15}$$

where $D^{ij}Q(\lambda|\lambda)$ denotes that $Q(\lambda|\lambda)$ has been differentiated i times with respect to the first argument and j times with respect to the second argument, are sufficient for (12) provided $d_j^T i(\lambda) d_k = 0 \quad j \neq k$. The purpose of condition (14) is that it implies:

$$\frac{1}{n}D^{20}Q(\lambda_n|\lambda) \rightarrow^P -i(\lambda) \qquad \text{as } n \to \infty, \tag{16}$$

which sometimes is easier to verify directly.

As an example consider a (m, p) curved exponential family, then $L_{COM}(\lambda)$ may be written

$$\sum_{i=1}^{m} \phi_i(\lambda) t_i(y_{COM}) - k(\phi_1(\lambda), .., \phi_m(\lambda)) + h(y_{COM}).$$

As is well known, see e.g Barndorff-Nielsen and Cox (1994) equation 2.120, $\frac{1}{n}D_{\lambda\lambda}L_{COM}(\lambda)$ has the form

$$\frac{1}{n}\sum_{i=1}^{m}(t_i - \eta_i)\frac{\partial^2 \phi_i}{\partial\lambda_r \partial\lambda_s} - i_{rs}(\lambda)$$

where $\eta_i = E_\lambda(T_i)$. This implies that the elements of $\frac{1}{n}D^{20}Q(\lambda_n|\lambda)$ can be written

$$\frac{1}{n}\sum_{i=1}^{m}[E_\lambda(T_i|Y_{OBS}) - \eta_i(\lambda_n)]\frac{\partial^2 \phi_i}{\partial\lambda_r \partial\lambda_s} - i_{rs}(\lambda_n), \tag{17}$$

where the conditional expectation is not a function of λ_n. Since typically $\eta_i(\lambda_n)$ converges to $E_\lambda(T_i)$, the question of whether (16) holds therefore boils down to whether the conditional expectations of $T_i \; i = 1, .., m$ converge to the unconditional as the number of observations increases. In models with repeated sampling the conditional expectation can often be expressed as a sum of terms depending on i. Hence due to the law of large numbers one can expect that the search vectors, $d_j \; j = 1, .., S$, are asymptotic I_{COM}-orthogonal provided $d_j^T i(\lambda) d_k = 0 \quad j \neq k$, under fairly general mechanisms describing the relation between the complete and observed data. In addition we see from (17) that condition (15) is satisfied if for all i:

$$\frac{1}{n}|E_{\lambda_n}(T_i|Y_{OBS}) - E_\lambda(T_i|Y_{OBS})| \rightarrow^P 0 \qquad \text{as } n \to \infty. \tag{18}$$

The underlying 'theme' of the EM and related algorithms is simplicity. Thus if one has an ECM-algorithm where each CM-step has a closed form solution, but with a convergence rate slower than that of EM, then it is not reasonable to re-construct the CM-steps to obtain an ECM with a convergence rate approximately equal to that of EM, if the new CM-steps are considerably more involved and require numerical techniques when executed. However it may be the case that for one subset of the parameter-vector there exists closed form solutions, when the other parameters are held fixed, but not closed form for the remaining set of the parameter-vector. In this case it may be an idea to have one CM-step take care of the first group, and then search in I_{COM}-orthogonal directions spanning the remaining set in the other CM-steps. Thus one will have not lost the simplification of the easy CM-steps, while at the same time not sacrificing convergence speed. A simple example of this case is given in Example 3 below.

4 Examples

Here three examples of the situation discussed in the previous section are given. Each example illustrates the case when the parameters in different CM-groups are orthogonal, and thus the search vectors are, as pointed out earlier, the standard basis vectors or collections of these. This is done to simplify the presentation, and because this case is the most important one in practice. Example 1 sheds light on a well-known ECM example that is of great importance. Example 2 applies the ECM to a more recent time-series model, and discusses its performance. Example 3 illustrates what gain there might be in basing an ECM-algorithm on I_{COM}-orthogonal directions, as opposed to not doing so.

Example 1: A multivariate normal regression model with incomplete data.

MR use this example, among two others, to motivate the ECM. Suppose the complete data consists of n independent observations from the k-dimensional model:

$$Y_i \sim N(X_i\beta, \Sigma), \tag{19}$$

where X_i is the design matrix $(k \times p)$ of the i-th observation, β a $(p \times 1)$ vector of unknown regression coefficients, and Σ a $(k \times k)$ unknown covariance matrix. MR point out that by specifying different structures on β and Σ, many important complete data models come out as special cases of (19), such as general repeated measures, see Jennrich and Schluchter (1986), and seemingly unrelated regressions, see Zellner (1962).

The maximum likelihood estimation of $\lambda = (\beta, \Sigma)$ is generally not in closed form, but observing that if either the mean vector or the covariance matrix where known, closed form solutions would exist, MR therefore define the following ECM-algorithm (for simplicity Σ is assumed unstructured):

E-step

Compute the conditional expectation of the complete data sufficient statistics, i.e. $E(Y_i|Y_{OBS}; \beta^{(t)}, \Sigma^{(t)})$ and $E(Y_iY_i^T|Y_{OBS}; \beta^{(t)}, \Sigma^{(t)})$.

CM-steps

1:

$$\beta^{(t+1)} = (\sum_{i=1}^{n} X_i^T(\Sigma^{(t)})^{-1}X_i)^{-1}(\sum_{i=1}^{n} X_i^T(\Sigma^{(t)})^{-1}Y_i),$$

2:

$$\Sigma^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\beta^{(t+1)})(Y_i - X_i\beta^{(t+1)})^T.$$

This is a considerably simpler algorithm than what an EM-algorithm applied to this model would be. And the interesting fact here is that because β and Σ are orthogonal (see Barndorff-Nielsen and Cox (1994) p.50), we have, by Proposition 2, that this added simplicity is in large samples 'free of charge'. This was not noted by MR.

To see how this works in practice, we have simulated data from the following bivariate version of (19):

$$(Y_{i1}, Y_{i2})^T \sim N(\begin{pmatrix}\beta\\ \beta\end{pmatrix}, \begin{pmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{pmatrix}).$$

Two sets of parameter values where used in the simulations. The first parameter set was $\lambda = (\beta, \sigma_1^2, \sigma_2^2) = (0, 1, 1)$, and the second parameter set was $\lambda = (0, 1, 1.25)$. Every Y_{ij} larger than 1 was censored. That $d_1^T \frac{1}{n} I_{COM}^{(n)} d_2 \to^P 0$, where $d_1 = (1, 0, 0)^T$ and $d_2 = [(0, 0, 1)^T, (0, 1, 0)^T]$, follows from $d_1^T i(\lambda) d_2 = 0$ and that $\frac{1}{n} I_{COM}^{(n)} \to^P i(\lambda)$ which in this example follows from the law of large numbers and the continuity of $E(Y_{ij}|Y_{ij} > 1; \lambda)$ as a function of λ, thus (13) holds here. The EM and ECM were applied to each simulated data set. The first data set had length 30, and the lengths were increased with increments of 60 to see what happens to $E|r_{ECM} - r_{EM}|$. (Each point on the following plot is the average of 20 values of the absolute value of $(r_{ECM} - r_{EM})$).

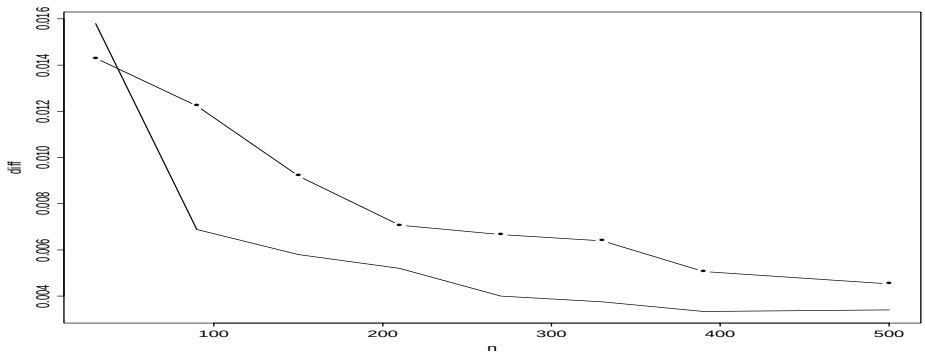

Figure 1: The figure shows $\frac{1}{20}\sum_{i=1}^{20} |r_{ECM}^{(i)} - r_{EM}^{(i)}|$ for increasing n. The solid line corresponds to: $\sigma_2^2 = 1$, and the dotted line to: $\sigma_2^2 = 1.25$.

Note that the graph from the parameter set resulting in relatively more censored values, the second set, converges quicker towards zero. This seems to be a general phenomenon, resulting from the increased number of variables being replaced by their conditional expectations.

Example 2: A hidden Markov autoregressive time series model.

This model was introduced by Hamilton(1989) to model economic time series with piece-wise constant mean and covariance-structures. Here we consider an AR(1) version of this model, but the conclusion, regarding the structure and behavior of the ECM-algorithm, apply equally well to the general model.

Suppose the time series $(Y_j)_{j=1}^{\infty}$ is generated by the model:

$$Y_j - \mu_{s_j} = \phi(Y_{j-1} - \mu_{s_{j-1}}) + \epsilon_j,$$

where $(s_j)_{j=1}^{\infty}$ is a two state Markov chain such that $p(s_j = i/s_{j-1} = i) = p_i$ for $i = 1, 2$, and $(\epsilon_j)_{j=1}^{\infty}$ is a sequence of independent $N(0, \sigma)$ variables. In the following we have conditioned on y_1 and knowledge of s_1, the unconditional likelihood is given in Hamilton (1993). In this model the level of the series at time 'j' is μ_{s_j}, and thus the state transitions of the Markov chain show up as level jumps in the time series. The Markov chain itself is however not observed. The general version of this model allows for higher order autoregressive behavior, that the autoregressive parameters can also shift with the Markov chain, and that the Markov chain can have more than 2 states with differing transition probabilities. For the remainder of this example we assume that $p_1 = p_2 = 0.5$, $\sigma = 1$ and $\mu_1 = 0$ are known, so that the unknown parameters are μ_2 and ϕ, i.e $\lambda = (\mu_2, \phi)$. Furthermore $Y_0 = 0$ and $s_0 = 1$

The maximum likelihood estimation for this model does not have closed form solutions, but as with the preceeding example if the mean parameters are known we can analytically solve for the autoregressive parameter, and vice versa. This leads to the following ECM-algorithm.

E-step

To find the $Q(|\lambda^{(t)})$ function here we need to calculate the so-called smoothed transition probabilities (see Hamilton(1993)), i.e $p(s_{j-1} = i, s_j = k|Y_{OBS}; \lambda^{(t)}) = p_j(i, k)$ for $j = 1, .., n$ and $k, i = 1, 2$.

CM-steps

1:

$$\mu_2^{(t+1)} = \frac{\sum_{j=2}^{n}(y_j - \phi^{(t)} y_{j-1})(p_j(2,1) - \phi^{(t)} p_j(1,2) + (1 - \phi^{(t)})p_j(2,2))}{\sum_{j=2}^{n}((\phi^{(t)})^2 p_j(1,2) + p_j(2,1) + (1 - \phi^{(t)})^2 p_j(2,2))},$$

2:

$$\phi^{(t+1)} = \frac{\sum_{j=2}^{n}\sum_{i,k=1}^{2}(y_{j-1} - \mu_{s_{j-1}}^{(t+1)})(y_j - \mu_{s_j}^{(t+1)})p_j(k,i)}{\sum_{j=2}^{n}\sum_{i=1}^{2}(y_{j-1} - \mu_{s_{j-1}}^{(t+1)})^2 p_j(i)}.$$

A special characteristic of this algorithm is that the E-step requires considerably more computer time than the two CM-steps. In an attempt to reduce the number of times the E-step is evaluated, one might be lead to iterate the CM-steps several times in-between each E-step, which also yields an ECM algorithm. While this strategy may work well in other models, since ϕ and μ here are orthogonal, iterating the CM-steps will not lead to large increases in $Q(|\lambda^{(t)})$, because in large samples executing the two CM-steps will in practice optimize this function. It is straightforward to show the orthogonality of ϕ and μ. To verify condition (12) in Proposition 2 here, i.e that $\frac{1}{n} d_1^T I_{COM}^{(n)} d_2 \to^P 0$, where $d_1 = (1, 0)^T$ and $d_2 = (0, 1)^T$, is more difficult than in the previous example. Noting that this is a (8,2) curved exponential model what has to be shown are conditions (17) and (18). However due to the complicated structure of the smoothed transition probabilities, $p(s_{j-1} = i, s_j = k|Y_{OBS}; \lambda)$, this is not attempted here, though it should be true under fairly weak assumptions.

To illustrate, we have simulated series under two different values of ϕ namely $\phi = 0.7$ and $\phi = -0.7$ while $\mu_2 = 3$ in each series. On each series the parameters were estimated with the above algorithm, call it ECM_1, and with an algorithm that iterates the CM-steps 50 times in-between each E-step, denote this algorithm by ECM_{50}. For each series length, 100 series

were simulated, and the average value of $|r_{ECM_1} - r_{ECM_{50}}|$ was calculated. The first series length is $n = 30$ and n is then increased with increments of 60. In Figure 2 the results are plotted.

We see that the negative correlation between successive values in the series makes $E|r_{ECM_1} - r_{ECM_{50}}|$ approach zero quicker, which is not surprising. Note also that the numerical value of $E|r_{ECM_1} - r_{ECM_{50}}|$ is small for all series lengths.

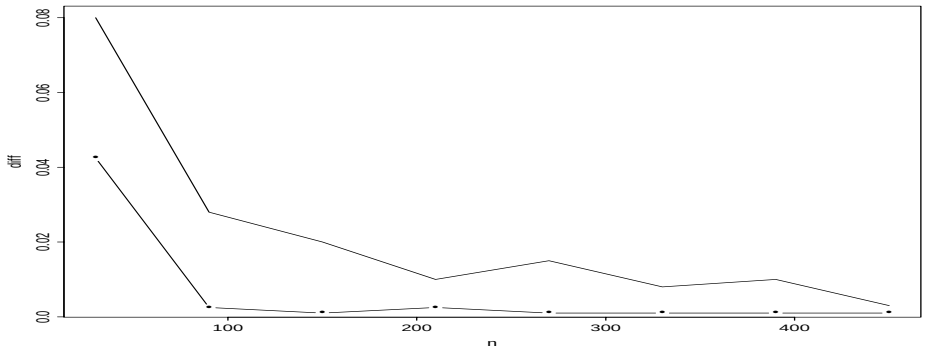

Figure 2: The figure shows $\frac{1}{100}\sum_{i=1}^{100}|r_{ECM_1}^{(i)} - r_{ECM_{50}}^{(i)}|$ for increasing n. The solid line corresponds to: $\phi = 0.7$, and the dotted line to: $\phi = -0.7$.

Example 3: A gamma model with incomplete data.

If the parameters that naturally belong to different CM-steps are not orthogonal, then it may be possible to reparameterize the model to obtain this property (Barndorff-Nielsen and Cox (1994) describe how to perform such an orthogonalization). We illustrate this idea in the following example, and present some simulation results that suggest that this can be quite effective.

This example was also used by MR to motivate the ECM-algorithm.

Here the complete data is a random sample from the gamma density:

$$f(y;\alpha,\beta) = \frac{y^{\alpha-1}\exp(-y/\beta)}{\beta^{\alpha}\Gamma(\alpha)}. \tag{20}$$

The ECM-algorithm presented by MR is:

E-step

Compute $z_i = E(y_i|Y_{OBS};\alpha^{(t)},\beta^{(t)})$ and $\log(z_i) = E(\log(y_i)|Y_{OBS};\alpha^{(t)},\beta^{(t)})$ for $i = 1,..,n$.

CM-steps

1:

$$\beta^{(t+1)} = \frac{\frac{1}{n}\sum_{i=1}^{n} z_i}{\alpha^{(t)}}, \tag{21}$$

2:

$$\alpha^{(t+1)} = \Psi^{-1}(\frac{1}{n}\sum_{i=1}^{n}\log(z_i) - \log(\beta^{(t+1)})),$$

where $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$. The second CM-step is solved e.g. by a 1-dim Newton-Raphson.

In dealing with complete data it is not difficult to show that the large sample convergence rate of the CM-algorithm, i.e. the ECM-algorithm without the E-step, is:

$$r_{CM} \approx \frac{1}{\alpha \Psi'(\alpha)}, \tag{22}$$

This is a monotonically increasing function of α which indicates that for larger α values the above ECM-algorithm may converge substantially slower than the corresponding EM. For $\alpha = 2$ $\frac{1}{\alpha \Psi'(\alpha)}|_{\alpha=2} = 0.76$.

One way we can try to increase the speed of convergence of this algorithm, is to replace the second CM-step in the above algorithm with a step that searches a maximum of $Q(|\lambda^{(t)})$, $\lambda = (\beta, \alpha)$ along the vector $d_2 = (a, 1)$ passing through the point $(\beta^{(t+1)}, \alpha^{(t)})$, where:

$$a = \frac{-\frac{\partial^2}{\partial \alpha \partial \beta} Q(\lambda|\lambda)}{\frac{\partial^2}{\partial^2 \alpha} Q(\lambda|\lambda)}. \tag{23}$$

The vector d_2 is I_{COM}-orthogonal to the search vector in the first CM-step which is $d_1 = (1, 0)$. This algorithm will converge by Proposition 1, at the rate of an EM-algorithm applied to this model and, of course, the data set at hand. This can considerably increase in convergence speed, as illustrated below. Thus we have constructed an algorithm that maintains the simplicity of the first CM-step, and replaced the 1-dim numerical optimization of the second CM-step with another 1-dim numerical optimization, thereby maintaining stability, however without loss of convergence speed compared to EM.

In this model it is, however, also possible to orthogonalize the parameters. By keeping α, we are lead to the new parameter $\theta = \alpha\beta$ which is orthogonal to α. The reparameterized density is:

$$f(y; \alpha, \theta) = \frac{y^{\alpha-1} \exp(-y\alpha/\theta)}{(\theta/\alpha)^\alpha \Gamma(\alpha)}. \tag{24}$$

It turns out that not only is θ orthogonal to α, the value of θ that maximizes the likelihood with respect to this parameter does not vary with α. The resulting ECM algorithm is therefore also an EM algorithm in the sense that executing the two CM-steps maximizes $Q(|\lambda^{(t)})$, and a trivial example of the result in Proposition 1 with $d_1 = (1, 0)^T$ and $d_2 = (0, 1)^T$ being the I_{COM}-orthogonal search vectors.

E-step

(Same as before)

CM-steps

1)

$$\theta^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} z_i,$$

2)

$\alpha^{(t+1)}$ is determined as the solution of:

$$\frac{1}{n} \sum_{i=1}^{n} \log(z_i) + \log(\alpha) - \alpha \log(\theta^{(t+1)}) - \frac{d}{d\alpha} \Gamma(\alpha) = 0.$$

We have simulated some data-sets from this gamma model and applied both of the algorithms. The number of simulated observations was in each case equal to 100, and every value larger

than unity was censored. In Table 1 the results are shown. Note that we have only varied the value of α bearing in mind (22), and fixated $\beta = 1$. Each entry in Table 1 is the average of 30 simulations.

Table 1: The table shows how convergence rates are effected by increases in α. $r_{ECM}^{(reg)}$ is the rate of the ECM based on (20) , and $r_{ECM}^{(ortho)}$ is the rate of the ECM based on (24).

α	1	2	3	4
$r_{ECM}^{(reg)}$	.60	.79	.87	.92
$r_{ECM}^{(ortho)}$	.02	.07	.19	.43

The table shows clearly that the orthogonalization has speed up the algorithm. Note also that since $ECM^{(ortho)}$ is also an EM algorithm, $r_{ECM}^{(ortho)}$ also gives the rates of the ECM where the first CM-step is as in (21) and the second CM-step searches along $d_2 = (a, 1)$, with a same as in (23), and the parameterization is as in (20).

5 Possible implications for ECME

Lui and Rubin (1994) introduced a related algorithm to the ECM, called the ECME algorithm. Here the idea is to try to increase the speed of ECM by replacing some of the (typically more difficult and lower dimensional) CM-steps with steps that conditionally maximize L_{OBS}. They present 3 compelling examples where ECME considerably outperforms EM, and thus presumably also ECM, both in convergence rate and number of iterations. There is however reason to believe that in some situations the EM will still outperform the ECME. Example 3 in the previous section illustrates that if there is a considerable amount correlation between the parameter estimators of parameters corresponding different CM-steps of the ECM, this algorithm can be considerably slower than the EM. Analogously one might expect that if the parameters in the CM steps that conditionally maximize L_{OBS} are strongly correlated to the parameters in the other CM-steps then ECME may be slower than EM, at least when the proportion of missing data is small. Let us illustrate this in the two-parameter situation. Suppose the model at hand has the parameters $\lambda = (\lambda_1, \lambda_2)$, where λ_1 and λ_2 are both scalar. Then it is not difficult to show, using the results of Lui and Rubin (1994), that the rate of ECME, r_{ECME}, when the Q-function, referring to (1), is maxized conditionally over λ_1 and L_{OBS} over λ_2, is:

$$r_{ECME} = \frac{(\frac{\partial^2}{\partial\lambda_1\partial\lambda_2}L_{OBS}(\lambda))^2}{\frac{\partial^2}{\partial\lambda_1^2}Q(\lambda|\lambda)\frac{\partial^2}{\partial\lambda_2^2}L_{OBS}(\lambda)} + \frac{\frac{\partial^2}{\partial\lambda_1^2}H(\lambda|\lambda)}{\frac{\partial^2}{\partial\lambda_1^2}Q(\lambda|\lambda)}, \tag{25}$$

where $Q(\lambda|\lambda)$ refers to (1) and $H(\lambda|\lambda) = Q(\lambda|\lambda) - L_{OBS}(\lambda)$. Thus if $|\frac{\partial^2}{\partial\lambda_1\partial\lambda_2}L_{OBS}(\lambda)|$ is relatively large, ECME might slow. We now illustrate the above ideas on data generated from a negative binomial model. The example will show that ECME can be considerably slower than EM, but after orthogonalizing the parameters in the different CM steps, as in Example 3, ECME is made considerably faster than it was and appreciably so compared to EM. We orthogonalize the parameters with respect to L_{COM}, which does not imply that the parameters are orthogonal with respect to L_{OBS}, but it is reasonable that this reduces $|\frac{\partial^2}{\partial\lambda_1\partial\lambda_2}L_{OBS}(\lambda)|$, and thus also (25).

Example 4: A negative binomial model with censored data.
Here the independent observations, say Y_i for $i = 1, .., n$, are generated from the following negative binomial distribution with density:

$$\frac{\Gamma(y+k)}{y!\Gamma(k)}\frac{\alpha^y}{(1+\alpha)^{y+k}} \qquad y = 0, 1, 2... \tag{26}$$

In this model the maximum likelihood estimators of the parameters k and α are considerably correlated. Introducing the new parameter $\nu = k\alpha$, gives a density of the form:

$$\frac{\Gamma(y+k)}{y!\Gamma(k)}\frac{\nu^y k^k}{(k+\nu)^{y+k}} \qquad y = 0, 1, 2..., \tag{27}$$

but now k and ν are orthogonal.

In the following we have simulated data sets from the above distribution and censored every variate larger than a constant c. We then applied to these data sets an EM-algorithm, an ECME-algorithm, call it $ECME_1$, based on the model formulation in (26), and an ECME-algorithm, call it $ECME_2$, based on the model formulation in (27). The $ECME_1$ maximizes at the t-th iteration $Q(\lambda|\lambda^{(t)})$, referring to (1), over α to obtain $\alpha^{(t+1)}$ conditionally on $k = k^{(t)}$, and then maximizes $L_{OBS}(\lambda)$ over k to find $k^{(t+1)}$ conditionally on $\alpha = \alpha^{(t+1)}$. The $ECME_2$ does the same as $ECME_1$, but now with α replaced by ν. The EM is by virtue of (1) and (3) already defined. One important note concerning the implementation of the EM is that the conditional expectations are not all in closed form. We can simplify, using properties of the gamma function:

$$\frac{\Gamma(y+k)}{y!\Gamma(k)} = \prod_{j=0}^{y-1}(k+j).$$

However evaluating $E(\log(\prod_{j=0}^{Y-1}(k+j))|y > c)$, and the derivatives of this expression, must be done numerically not only every E-step, but also in every iteration of the numerical optimizing routine in the M-step, where we used a Newton-Raphson. This is a serious drawback and a strong argument for using the ECME-algorithms that circumvent this problem since it is not necessary to evaluate $E(\log(\prod_{j=0}^{Y-1}(k+j))|y > c)$ in order to maximize $Q(\lambda|\lambda^{(t)})$ over α. We do however have that:

$$E(Y|Y > c; \nu, k) = \frac{\nu - \sum_{y=0}^{c-1}(\frac{k}{\nu})^k(\frac{1}{\frac{k}{\nu}+1})^{y+k}\prod_{j=0}^{y-1}(j+k)}{1 - F(Y \leq c; \nu, k)},$$

using the fact that the negative binomial is a mixture of a gamma and a Poisson distribution and changing the orders of summation and integration.

In the following table we show the rates of the 3 algorithms when applied to 4 simulated data sets, each of length 1000 and generated under a different value of α. For each sample there has been used a different value of α in the parameterization in (26), as indicated in the table, while keeping $k = 10$. Every generated variate larger than $c = 5$ has been censored.

Table 2: The table shows the rates of convergence of different algorithms on negative binomial model.

α	r_{EM}	r_{ECME_1}	r_{ECME_2}
.35	.38	.99	.11
.45	.54	.99	.25
.55	.70	.99	.48
.65	.82	.99	.63

The table shows that the $ECME_1$ algorithm is very slow, considerably slower than EM, while this difference in rate decreases when the amount of censored variates increases, corresponding increases in α. The $ECME_2$ algorithm converges however appreciably faster than EM.

6 Concluding remarks

We have shown that it is possible, however not always desirable, to construct an ECM-algorithm to converge at the same or approximately the same rate as EM. This gave insight into the performance of the ECM in some practically useful models, and suggested possible ways to speed up its convergence. It also illustrated the importance of parameter orthogonality for computational purposes, as noted by Cox and Reid (1987). The third example demonstrated that an algorithm based on what we called I_{COM}-orthogonal search vectors can lead to substantially quicker convergence than an ECM not constructed in this manner. The advantages of an ECM with I_{COM}-orthogonal search directions over that of an EM algorithm, assuming that this algorithm does not have a closed form M-step, is that the dimension of the numerical optimization is reduced, thus increasing the stability of the algorithm without sacrificing convergence speed, at least in the quadratic region close to the maximal point . The disadvantage is that an ECM thusly constructed may be more tedious to implement. The possible advantages of the type of ECM we discuss , is that it may converge at a quicker rate, than in other implementations, however the disadvantages may be that it requires more effort to implement and that the CM-steps here may take longer time to evaluate. The importance of reducing the correlation between the parameters in the CM-steps of an ECME-algorithm was also illustrated.

References

[1] Cox, D. R. and Reid, N., (1987) " Parameter orthogonality and approximate conditional inference" *J. Roy. Statist. Soc. Ser. B* 49: 1-39.

[2] Barndorff-Nielsen, O. E. and Cox, D. R., (1994) "Inference and Asymptotics", Chapman and Hall.

[3] Dempster, A. P., Laird, N. M. and Rubin, D. B., (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm", *J. Roy. Statist. Soc. Ser. B* 39: 1-38.

[4] Hamilton, J. D. (1989), " A new approach to the analysis of nonstationary time series and the business cycle", *Econometrica* 57, 357-384.

[5] Hamilton, J. D. (1993), "Handbook of Statistics", Vol. 11, Elsevier Science Publishers B.V.

[6] Jennrich, R. I. and Schluchter, M. D. (1986), "Unbalanced repeated measures models with structured covariance matrices.", *Biometrics*, 42, 805-820.

[7] Luenberger, D. G., (1989), "Linear and Nonlinear Programming", Wesley.

[8] Lui, C. and Rubin, D. B., (1994), "The ECME algorithm: A simple extension of the EM and ECM with faster monotone convergence", *Biometrika*, 81, 4, 633-648.

[9] Meng, X. L. and van Dyke, D., (1997), "The EM algorithm- an Old Folk-song sung to a fast new tune" *J. Roy. Statist. Soc. Ser. B* 59: 511-567.

[10] Meng, X. L. and Rubin, D. B., (1991), "Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm", *J. Am. Statist. Assoc.* 86, 899-910.

[11] Meng, X. L. and Rubin, D. B., (1992), "Recent Extensions to the EM Algorithm", In *Bayesian Statistics 4*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 307-320, Oxford University Press.

[12] Meng, X. L. and Rubin, D. B., (1993), "Maximum likelihood estimation via the ECM algorithm: A general framework", *Biometrika* 80, 2, 267-278.

[13] Meng, X. L., (1994), "On the rates of convergence of the ECM algorithm", *The Annals of Statistics*, Vol. 22, No. 1, 326-339.

[14] Wu, C. F. J., (1983), "On the convergence properties of the EM algorithm", *The Annals of Statistics*, Vol. 11, 95-103.

[15] Zangwill, W. (1969), "Nonlinear Programming-A Unified Approach", Prentice-Hall, Englewood Cliffs, NJ.

[16] Zellner, A. (1962), "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias." *J. Am. Statist. Assoc.* 57, 348-368.